Topic Modeling

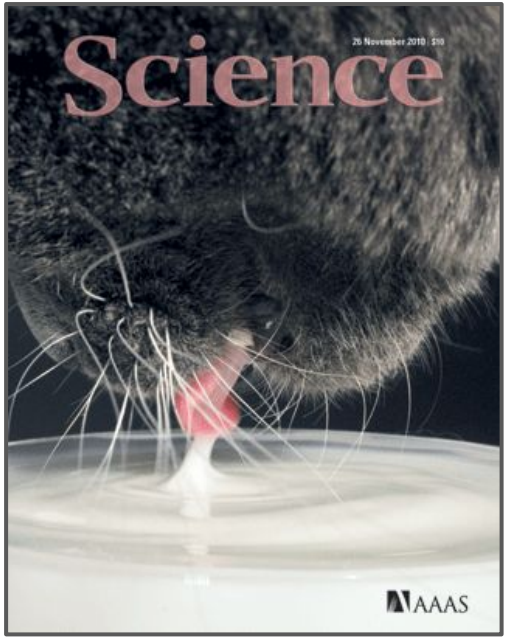
Chris Murdock

Why Science?

- Peer-reviewed journal and one of the world’s top academic journals
    - Less than 7% of submitted papers are published

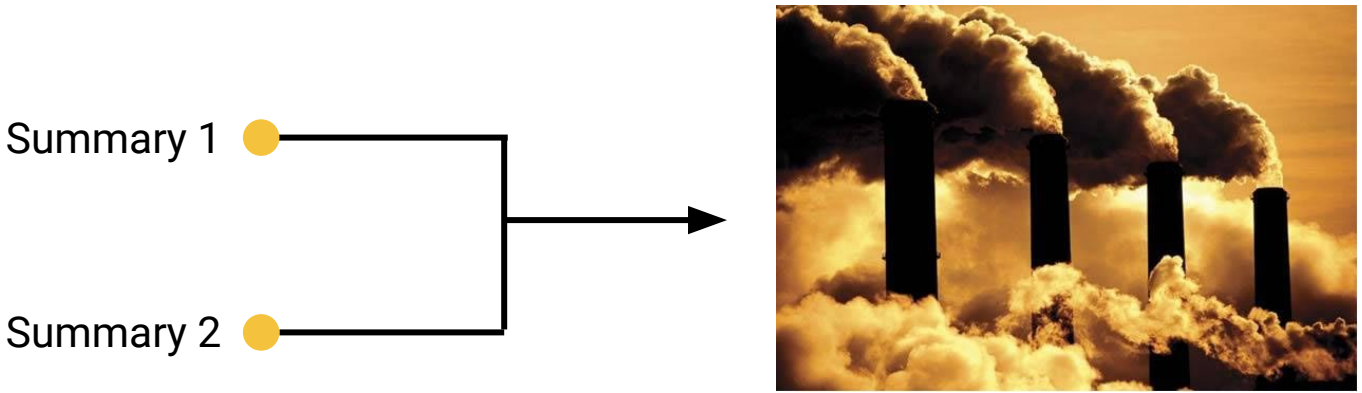
Why Science?

- Peer-reviewed journal and one of the world's top academic journals
- Unlike most scientific journals, Science covers all scientific disciplines



Objective

Can we use summaries to find topics or overarching themes? If so, what information can we take away about how topics trend over time?



Global Warming

Analysis Pipeline

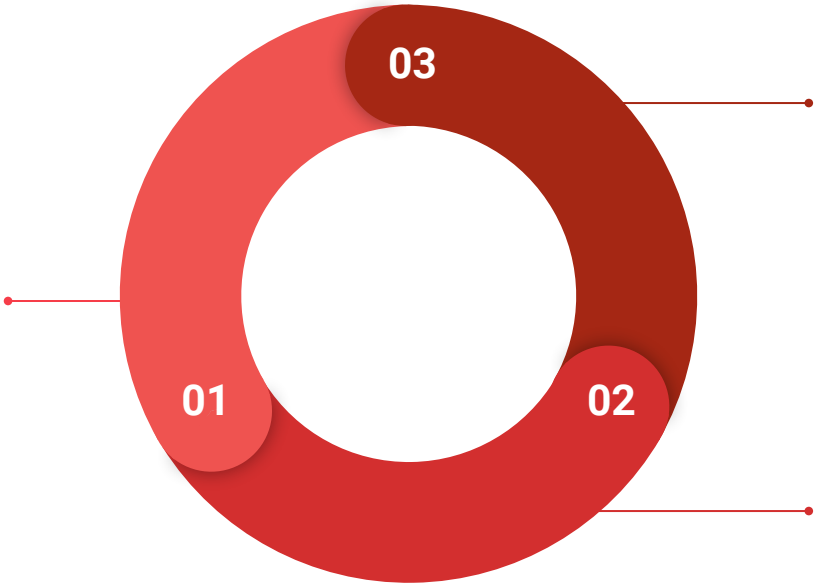
Scrape articles from Science's website from 2001-present (~35,000 articles)

- Article summaries, titles and year published were loaded into MongoDB

Analysis Pipeline

Preprocess Text

Text cleaning (removing stop words and punctuation).



Topic Modeling

NMF model

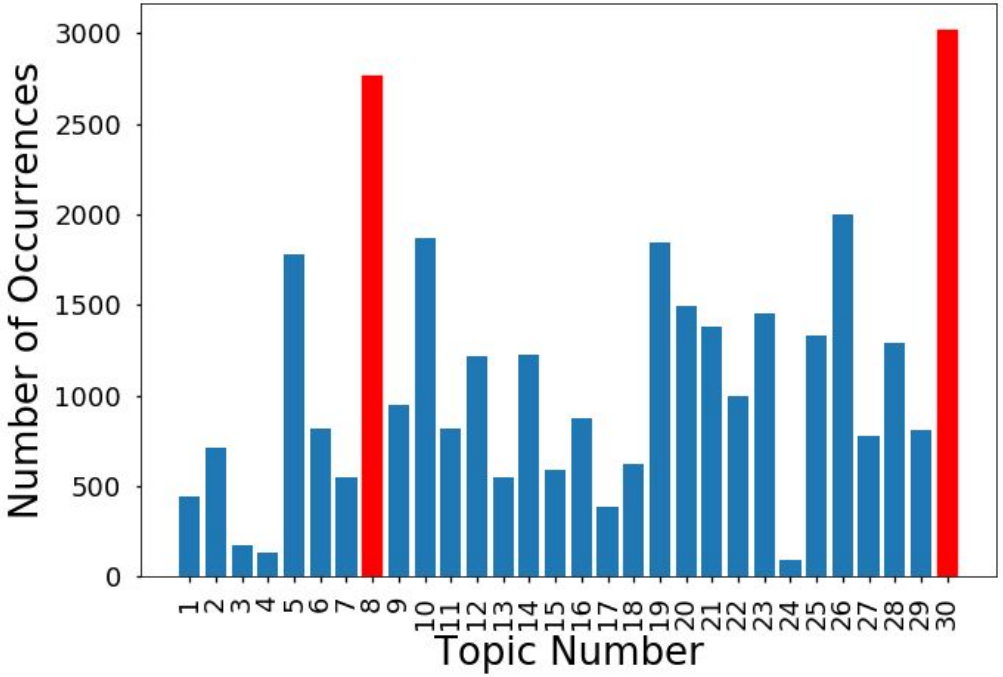
Vectorize Text

Tf-idf Vectorizer

Topics from 2001-2016

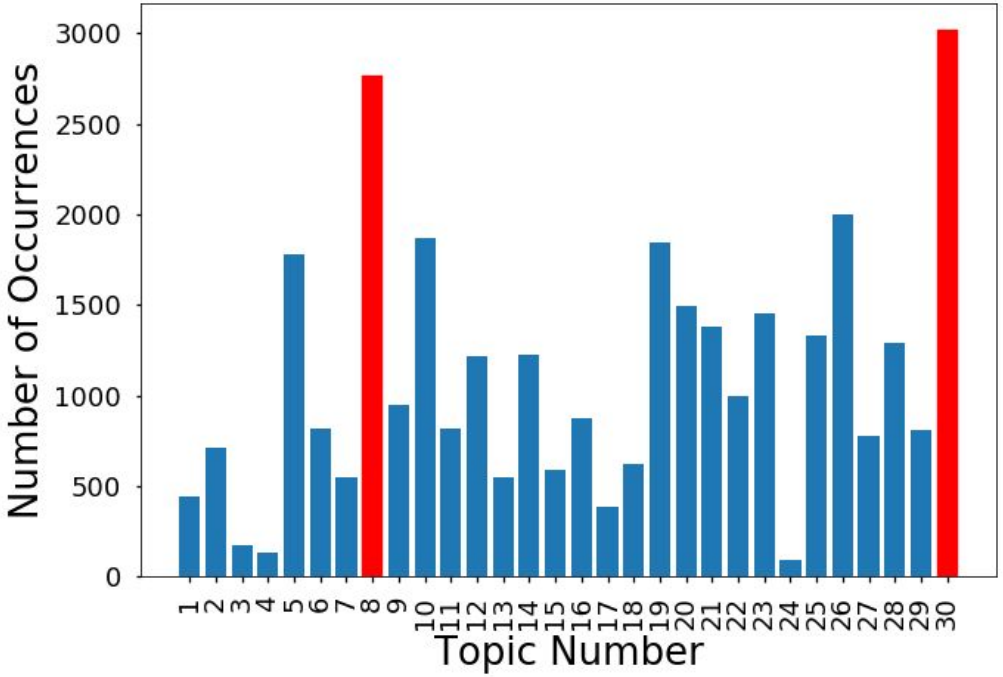
- 33,000 articles narrowed down to 30 topics
- Each document assigned to a topic

Topics from 2001-2016



- 33,000 articles narrowed down to 30 topics
- Each document assigned to a topic

Topics from 2001-2016

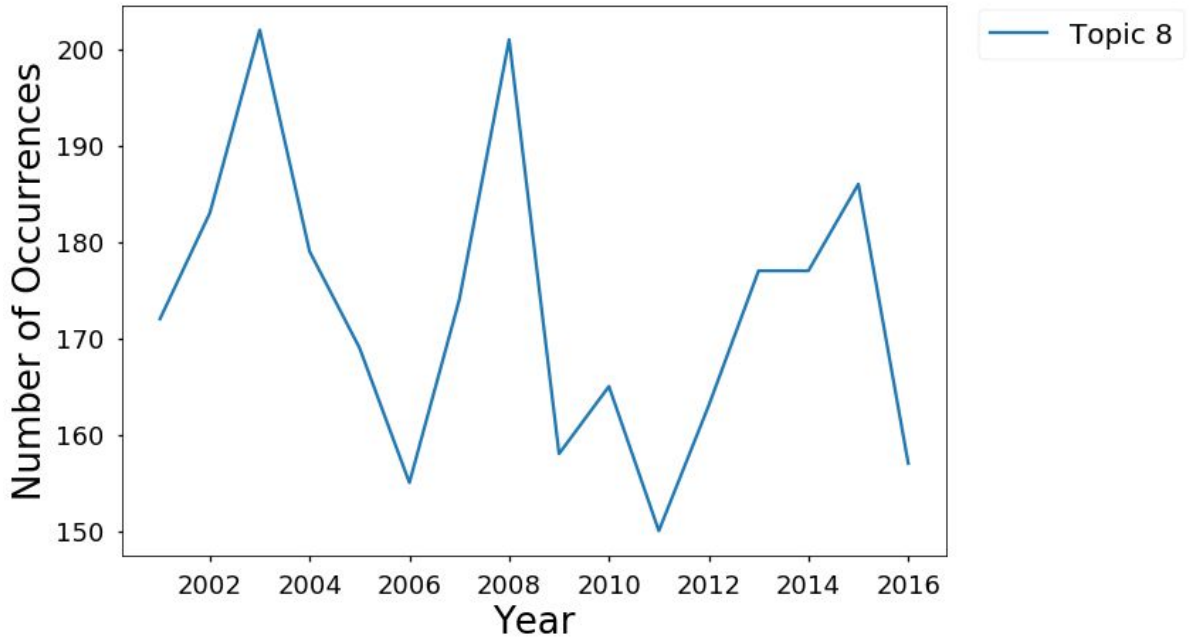


Main Topics:

8 - material electron structure magnetic use field temperature surface high property

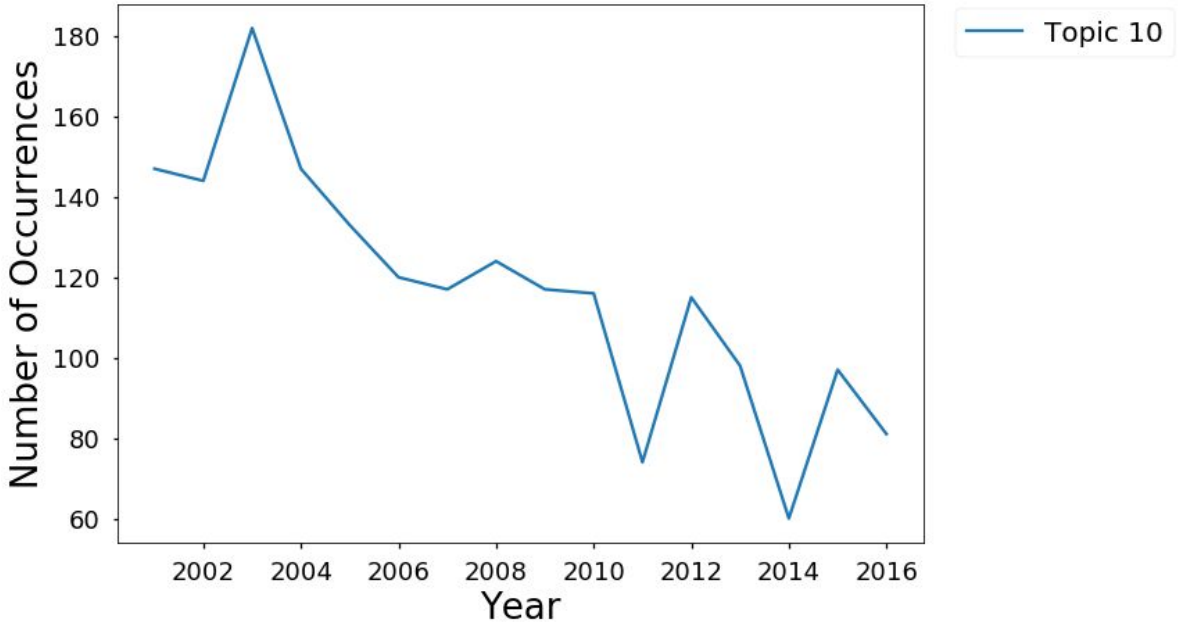
30 - state world united united state country health china nation people public

Trends in Topics



material electron structure magnetic use field temperature surface high property

Downward Trends



protein membrane signal bind receptor complex structure domain pathway kinase

Future Steps

- Collect more articles to look at topic trends over decades.

- Look at topics for a given year
  - 2017 main topic:
    - science research scientist u.s world trump president new government agency